Published in final edited form as:
Stat Interface. 2014 July 1; 6(3): 315–324. doi:10.4310/SII.2013.v6.n3.a2.

A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models

Qingxia Chen[†] and
Department of Biostatistics, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, 3723, USA

Joseph G. Ibrahim[‡]
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

Qingxia Chen: cindy.chen@vanderbilt.edu; Joseph G. Ibrahim: ibrahim@bios.unc.edu

Abstract

Multiple Imputation, Maximum Likelihood and Fully Bayesian methods are the three most commonly used model-based approaches in missing data problems. Although it is easy to show that when the responses are missing at random (MAR), the complete case analysis is unbiased and efficient, the aforementioned methods are still commonly used in practice for this setting. To examine the performance of and relationships between these three methods in this setting, we derive and investigate small sample and asymptotic expressions of the estimates and standard errors, and fully examine how these estimates are related for the three approaches in the linear regression model when the responses are MAR. We show that when the responses are MAR in the linear model, the estimates of the regression coefficients using these three methods are asymptotically equivalent to the complete case estimates under general conditions. One simulation and a real data set from a liver cancer clinical trial are given to compare the properties of these methods when the responses are MAR.

Keywords and phrases

Missing data; Multiple imputation; Maximum likelihood; Fully Bayesian; Missing response; Missing at random

1. INTRODUCTION

Missing data is very common in various experimental settings, including clinical trials, sample surveys and environmental studies. There are essentially three major likelihood-based approaches for handling missing data in a regression problem. These are i) Maximum Likelihood (ML), ii) Multiple Imputation (MI), and iii) Fully Bayesian (FB). The EM

Correspondence to: Qingxia Chen, cindy.chen@vanderbilt.edu.

[†]Dr. Chen's work was partially supported by 1R21HL097334 and UL1 RR024975-01 from the National Institutes of Health.
[‡]Dr. Ibrahim's work was partially supported by CA 70415 and GM 70335 from the National Institutes of Health.

algorithm is a technique often used to obtain ML estimates and is useful when the likelihood function of the observed data has no closed form. The recent developments of missing data approaches also include empirical likelihood method [18], parametric fractional imputation [10], among others. In this paper, we investigate theoretical connections between MI, ML (especially within the EM framework), and FB approaches in the linear regression model when the response variable is missing at random (MAR).

It is well known that when the response variable is MAR and the covariates are fully observed, the likelihood function of the observed data is the same as the complete case likelihood (i.e. the likelihood obtained by omitting all cases with missing values), and therefore the ML estimates are identical to the complete case (CC) estimates. However, this result is not obvious under the MI and FB approaches. Although the CC estimates are unbiased and efficient under MAR responses, the MI and FB methods are still used in practice since many researchers are unaware of this special property of the CC estimates. To study the relationships between the three methods in this context, we consider MAR responses in the linear model and investigate the small and large sample properties of the estimates, and derive analytic and asymptotic expressions of the estimates and standard errors for the MI, ML and FB approaches. Under noninformative priors for the MI and FB methods, we show that the estimates and their standard errors under these three approaches are asymptotically equivalent to the CC estimates.

There is much literature on MI, ML and FB, respectively. [21] provide asymptotic results for MI with MAR responses in linear models. [17], [15], and [20] discuss theoretical properties of proper and improper MI. [22] and [19] propose a consistent variance estimator for MI. For ML, one of the earliest references is [12]. [6], [11], and [8] proposes the “EM by the method of weights” and the Monte Carlo EM algorithm (MCEM) for the ML framework in generalized linear models (GLM). [16] examine the problem of using EM to obtain the asymptotic covariance matrix of the parameter estimates. [7] discuss FB methods for MAR covariates in GLM’s. There are two major differences of our work in this paper from previous literature. First, for MAR responses, we derive both the small and large sample properties of the estimates, while the previous work mainly focuses on large sample properties. Secondly, the main purpose of this paper is to investigate the theoretical relationships between MI, ML and FB, as this was only investigated only through simulations before.

The rest of this paper is organized as follows. In Section 2.1, we derive the small sample and asymptotic expressions of the estimates and standard errors for proper MI. In Section 2.2, we derive the expressions for ML via the EM algorithm. The expressions for FB are derived in Section 2.3. In Section 3, we conduct a simulation to demonstrate our results. A real data analysis of a liver cancer clinical trial is given in Section 4, and we conclude the paper with a brief discussion in Section 5.

2. MAR RESPONSES IN THE LINEAR MODEL

Consider the linear model

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{e}, \quad (1)$$

where β is a $p \times 1$ vector of unknown parameters, $\mathbf{X}$ is an $n \times p$ full rank matrix of explanatory variables including an intercept, and $\mathbf{e}$ is an $n \times 1$ vector of random errors with $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, where σ^2 is assumed unknown throughout. We assume throughout that $\mathbf{X}$ is fully observed and the components of $\mathbf{y}$ are MAR. For simplicity, we rearrange the data so that $\mathbf{y}_1 = (y_1, \ldots, y_{n_1})'$ are fully observed and $\mathbf{y}_2 = (y_{n_1+1}, \ldots, y_n)'$ are MAR, and assume that the corresponding $n_1 \times p$ and $n_2 \times p$ matrices of fixed covariates $\mathbf{X}_1$ and $\mathbf{X}_2$ for $\mathbf{y}_1$ and $\mathbf{y}_2$ are full-rank, $n_1+n_2 = n$ and $p < n_1$. Therefore, we write $\mathbf{y}=(\mathbf{y}_1', \mathbf{y}_2')'$ and $\mathbf{X}=(\mathbf{X}_1', \mathbf{X}_2')'$.

As shown in [13], the maximum likelihood estimates of β and σ^2 are the same as the CC estimates in the linear regression model, in which cases with any missing values are simply discarded. In fact, this is true for any regression model with MAR responses satisfying conditional independence between $\mathbf{y}_1$ and $\mathbf{y}_2$ given $\mathbf{X}$ and γ, where $\gamma = (\beta, \sigma^2)$, since the likelihood function of the observed data $\mathbf{D}_{obs} = (\mathbf{y}_1, \mathbf{X})$ is given by

$$\begin{aligned} L(\gamma|\mathbf{D}_{obs}) &= \int p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{X}, \gamma)\, d\mathbf{y}_2 \\ &= p(\mathbf{y}_1|\mathbf{X}, \gamma) \int p(\mathbf{y}_2|\mathbf{X}, \gamma)\, d\mathbf{y}_2 \\ &= p(\mathbf{y}_1|\mathbf{X}, \gamma), \end{aligned}$$

which is the CC likelihood.

The standard results for the linear regression model with MAR responses are

$$\hat{\beta}_{ML}=\hat{\beta}_{CC}=(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1, \quad (2)$$

which is independent of

$$\hat{\sigma}^2_{ML}=\hat{\sigma}^2_{CC}=\mathbf{y}_1'(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}_1/n_1 \quad (3)$$

with $E(\hat{\beta}) = \beta$ and $E(\hat{\sigma}^2_{ML})=(n_1-p)\sigma^2/n_1$. The variances of the estimates are

$$\mathrm{Var}(\hat{\beta}_{ML})=\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \quad (4)$$

and

$$\mathrm{Var}(\hat{\sigma}^2_{ML})=2(n_1-p)\sigma^4/n_1^2. \quad (5)$$

Clearly, we can adjust the estimate of σ^2 to get an unbiased estimate by letting $\tilde{\sigma}^2_{ML}=n_1\hat{\sigma}^2_{ML}/(n_1-p)$. It is worth noting that if we apply the EM algorithm in this setting and use Louis's method [14] to get the variance estimates, we get same variance estimates of

$\hat{\beta}_{ML}$ as in Eq. (4), and the variance estimate of $\hat{\sigma}^2_{ML}$ is equal to $2\hat{\sigma}^4_{ML}/n_1$, which is larger than Eq. (5) in small samples but asymptotically equivalent when $n_1 \to \infty$. In the next three subsections, we explore the small and large sample properties of the estimators under MI, ML via the EM algorithm, and FB methods for MAR responses under model (1).

2.1 Multiple imputation (MI)

Multiple imputation has emerged as a very popular technique for inference in missing data problems. In this section, we consider the precision parameter instead of the variance parameter for the development of MI. Therefore, we assume $\gamma^* = (\beta, \tau)$, where $\tau = 1/\sigma^2$. Proper MI is based on creating imputed datasets in which the missing values are sampled from the posterior predictive distribution of the missing data given the observed data, given by

$$p(\mathbf{D}_{mis}|\mathbf{D}_{obs}) = \int p(\mathbf{D}_{mis}|\mathbf{D}_{obs}, \gamma^*)\pi(\gamma^*|\mathbf{D}_{obs})\, d\gamma^*, \quad (6)$$

where $\mathbf{D}_{obs} = (\mathbf{y}_1, \mathbf{X}_1, \mathbf{X}_2)$, and $\mathbf{D}_{mis} = \mathbf{y}_2$ for the current setting. $\pi(\gamma^*|\mathbf{D}_{obs})$ is the posterior distribution of γ^* based on the observed data, given by $\pi(\gamma^*|\mathbf{D}_{obs}) \propto \{\int L(\gamma^*|\mathbf{D}_{mis}, \mathbf{D}_{obs})d\mathbf{D}_{mis}\}\pi(\gamma^*)$, where $L(\gamma^*|\mathbf{D}_{mis}, \mathbf{D}_{obs})$ is the likelihood function of the complete-data, $\int L(\gamma^*|\mathbf{D}_{mis}, \mathbf{D}_{obs})d\mathbf{D}_{mis}$ is the likelihood based on the observed data, and $\pi(\gamma^*)$ is the prior distribution of γ^*. Assume $\mathbf{D}^{(l)}_{mis}$, $l = 1, \ldots, m$, are draws of $\mathbf{D}_{mis}$ from the posterior predictive distribution $p(\mathbf{D}_{mis}|\mathbf{D}_{obs})$ given in Eq. (6). Let $\hat{\gamma}^*_l$ and $\mathbf{V}_l$ denote the posterior mean and covariance matrix of γ^* based on $\pi(\gamma^*|D_{obs})$ calculated for the lth imputed data set $(\mathbf{y}_1, \mathbf{y}^{(l)}_2)$. Then, the MI estimate of γ^* is $\hat{\gamma}^*_{MI} = m^{-1}\sum_{l=1}^m \hat{\gamma}^*_l$, and the estimate of the variance of $\hat{\gamma}^*_{MI}$ is

$$\hat{\mathrm{Var}}(\hat{\gamma}^*_{MI}) = \hat{\mathbf{V}} + \left(1 + \frac{1}{m}\right)\hat{\mathbf{B}}, \quad (7)$$

where $\hat{\mathbf{V}} = m^{-1}\sum_{l=1}^m \hat{\mathbf{V}}_l$ and $\hat{\mathbf{B}} = \sum_{l=1}^m (\hat{\gamma}^*_l - \hat{\gamma}^*_{MI})(\hat{\gamma}^*_l - \hat{\gamma}^*_{MI})'/(m-1)$ is the between-imputation variance. There are several imputation methods that have been proposed for the MI method. In this paper, we concentrate on proper MI using the improper prior,

$$\pi(\gamma^*) \propto \tau^{-1}. \quad (8)$$

We note that in MI, the imputation model can be different from the analysis model, but in this paper we only consider the case in which the two models are the same.

Theorem 1 gives the small sample behavior of the estimates of β and σ^2 for proper MI assuming the improper prior $\pi(\gamma^*) \propto \tau^{-1}$. Large sample properties of the estimates are also given under some general conditions. To derive these properties, we need the following lemma.

Lemma 2.1—If the $n\times 1$ random vector $\mathbf{z}$ has a multivariate t distribution, denoted $S_n(v, \mu, \mathbf{V})$, with density proportional to $[1+\frac{1}{v}(\mathbf{z}-\mu)'\mathbf{V}^{-1}(\mathbf{z}-\mu)]^{-\frac{(\mathbf{v}+\mathbf{n})}{\mathbf{2}}}$, and $\mathbf{A}$ and $\mathbf{B}$ are matrices of constants, then

1. $E(\mathbf{z}'\mathbf{A}\mathbf{z})=\frac{v}{v-2}tr(\mathbf{A}\mathbf{V})+\mu'\mathbf{A}\mu$, when $v > 2$
2. $E(\mathbf{z}\mathbf{z}'\mathbf{A}\mathbf{z})=\frac{v}{v-2}[\mathbf{V}\mathbf{A}'\mu+\mathbf{V}\mathbf{A}\mu+\mu tr(\mathbf{A}\mathbf{V})]+\mu\mu'\mathbf{A}\mu$, when $v > 2$
3. $$\begin{aligned}E[(\mathbf{z}'\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{B}\mathbf{z})] &= \frac{v^2}{(v-2)(v-4)}[tr(\mathbf{A}\mathbf{V})tr(\mathbf{B}\mathbf{V})+tr(\mathbf{A}\mathbf{V}\mathbf{B}\mathbf{V})+tr(\mathbf{A}\mathbf{V}\mathbf{B}'\mathbf{V})]\\ &+\frac{v}{v-2}[tr(\mathbf{A}\mathbf{V})\mu'\mathbf{B}\mu\\ &+\mu'\mathbf{A}\mu tr(\mathbf{B}\mathbf{V})\\ &+\mu'(\mathbf{A}\mathbf{V}\mathbf{B}\\ &\quad +\mathbf{A}'\mathbf{V}\mathbf{B}+\mathbf{A}\mathbf{V}\mathbf{B}'+\mathbf{A}'\mathbf{V}\mathbf{B}')\mu]\\ &+\mu'\mathbf{A}\mu\mu'\mathbf{B}\mu \quad \text{, when } v > 4.\end{aligned}$$

The proof of Lemma 2.1 is given in the Appendix. For the linear regression model (1) with prior as Eq. (8), the posterior distribution of γ^* based on the observed data is

$$\begin{aligned}p(\gamma^*|\mathbf{y}_1) &\propto p(\mathbf{y}_i|\gamma^*)\pi(\gamma^*)\\ &\propto \tau^{n_1/2-1}\exp\left\{-\frac{\tau}{2}(\mathbf{y}_1-\mathbf{X}_1\beta)'(\mathbf{y}_1-\mathbf{X}_1\beta)\right\}\end{aligned}$$

and the posterior predictive distribution is

$$\begin{aligned}p(\mathbf{y}_2|\mathbf{y}_1) &= \int\int p(\mathbf{y}_2|\mathbf{y}_1,\gamma^*)p(\gamma^*|\mathbf{y}_1)\,d\beta d\tau\\ &\propto \left[1+\frac{(\mathbf{y}_2-\hat{\mathbf{y}}_2)'\mathbf{H}(\mathbf{y}_2-\hat{\mathbf{y}}_2)}{(n_1-p)s^2}\right]^{-\frac{n_1+n_2-p}{2}}\end{aligned}$$

where $\hat{\mathbf{y}}_2=\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$, $s^2=\mathbf{y}_1'(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}_1/(n_1-p)$ and $\mathbf{H}=\mathbf{I}-\mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2'$. Since $\mathbf{X}'\mathbf{X}=\mathbf{X}_1'\mathbf{X}_1+\mathbf{X}_2'\mathbf{X}_2$, and $\mathbf{X}_1'\mathbf{X}_1$ are of full-rank, it can be shown that $\mathbf{H}$ is positive definite with inverse $\mathbf{H}^{-1}=\mathbf{I}+\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2'$. Hence, the posterior predictive distribution of $[\mathbf{y}_2|\mathbf{y}_1]$ is a multivariate t distribution given by

$$\mathbf{y}_2|\mathbf{y}_1\sim S_{n-n_1}(n_1-p, \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1, s^2\mathbf{H}^{-1}). \quad (9)$$

Theorem 2.1 establishes the small and large sample properties of the estimates based on the MI method.

Theorem 2.1—For the linear regression model (1) with prior (8), let $\mathbf{y}_2^{(l)}$, $l = 1, \ldots, m$, be the samples of $\mathbf{y}_2$ from $[\mathbf{y}_2|\mathbf{y}_1]$ in Eq. (9). Then

i. the multiple imputation (MI) estimate of β is

$$\hat{\beta}_{MI}=\frac{1}{m}\sum_{l=1}^{m}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_1'\mathbf{y}_1+\mathbf{X}_2'\mathbf{y}_2^{(l)}), \quad (10)$$

with mean $E(\hat{\beta_{MI}}) = \beta$ and variance

$$Var(\hat{\beta}_{MI})=\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}+\frac{\sigma^2(n_1-p)}{m(n_1-p-2)}[(\mathbf{X}_1'\mathbf{X}_1)^{-1}-(\mathbf{X}'\mathbf{X})^{-1}]. \quad (11)$$

ii. The MI estimate of σ^2 is

$$\hat{\sigma}^2_{MI}=\frac{1}{m}\sum_{l=1}^{m}y^{(l)'}(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}^{(l)}/(n-2), \quad (12)$$

with mean

$$E(\hat{\sigma}^2_{MI})=\frac{(n-p-2)(n_1-p)}{(n-2)(n_1-p-2)}\sigma^2$$

and variance

$$Var(\hat{\sigma}^2_{MI})=\frac{2(n_1-p)(n-p-2)^2}{(n-2)^2(n_1-p-2)^2}\sigma^4+a_1\sigma^4/m, \quad (13)$$

where $a_1=\frac{2(n-n_1)(n-p-2)(n_1-p)(n_1-p+2)}{(n-2)^2(n_1-p-2)^2(n_1-p-4)}$.

From Theorem 2.1, it can be shown that the MI estimate of β and σ^2 as well as their variances are asymptotically equivalent to the CC estimates. Furthermore, after some algebra, we can show that when $n > n_1 > p+ 2$

$$E(\hat{\sigma}^2_{MI}|\mathbf{y}_1)=\frac{n_1(n-p-2)}{(n-2)(n_1-p-2)}\hat{\sigma}^2_{CC}>\hat{\sigma}^2_{CC},$$

and

$$\mathrm{Var}(\hat{\sigma}^2_{MI})=\frac{n_1^2(n-p-2)^2}{(n_1-p-2)^2(n-2)^2}\mathrm{Var}(\hat{\sigma}^2_{CC})+\frac{2(n-p+2)(n-n_1)}{m(n_1-p+2)(n+2)}\sigma^4>\mathrm{Var}(\hat{\sigma}^2_{CC}),$$

where $\hat{\sigma}^2_{CC}$ is given in Eq. (3). We note here that throughout this paper, we do not consider the situation in which the number of regression coefficients *p* increases as *n* increases, so *p* is either fixed or increases at a slower rate than *n*.

Remark 2.1— $E(\hat{\sigma}^2_{MI})$ is independent from *m* while $\mathrm{Var}(\hat{\sigma}^2_{MI})$ is a function of *m*, therefore, increasing the number of imputations, *m*, does not reduce the bias of $\hat{\sigma}^2_{MI}$, but it reduces the variance of $\hat{\sigma}^2_{MI}$.

Remark 2.2— $(\hat{\beta_{MI}}|\mathbf{y}_1)/\hat{\beta_{CC}} \to 1$, $(\tilde{\sigma}^2_{MI}|\mathbf{y}_1)/\tilde{\sigma}^2_{CC} \to 1$ as $m \to \infty$, where $\tilde{\sigma}^2_{MI} = \frac{(n-2)(n_1-p-2)}{(n-p-2)(n_1-p)}\hat{\sigma}^2_{MI}$ and $\tilde{\sigma}^2_{CC} = \frac{n_1}{n_1-p}\hat{\sigma}^2_{CC}$ are unbiased estimates of σ^2. However, this is not true for a fixed *m*.

We also note that $\mathrm{Var}(\hat{\beta_{MI}}) > \mathrm{Var}(\hat{\beta_{CC}})$ and $\mathrm{Var}(\tilde{\sigma}^2_{MI}) > \mathrm{Var}(\tilde{\sigma}^2_{CC})$, which imply that the MI estimates are less efficient than the CC estimates. This is because the ML estimates for MAR responses in the linear model are the same as the CC estimates, and the ML estimates are most efficient if the model is correct. The extra variability of the MI estimate is induced by the sampling involved in finding the estimator. Even though we are able to improve the MI estimates under the setting of MAR responses in linear regression with small samples, this is not the main aim of this paper. The goal of this paper is to investigate the relationships between MI, ML, and FB approaches. The small sample properties of MI have been studied under more general settings in [1] and [9].

2.2 Maximum likelihood (ML)

As shown in equations (2) and (3), there are closed form estimates of β and σ^2 using the ML method when the response variable is MAR in the linear model, and those estimates are precisely the CC estimates. However, the ML method is more generally carried out using the EM algorithm, which can be either directly solved when the E-step has a closed form, or it may be obtained using Monte Carlo methods when it does not have a closed form. This latter version of the EM algorithm is referred to as the Monte Carlo EM (MCEM) algorithm and is a more general method of carrying out ML since for most regression models with missing data, the E-step does not have a closed form. We will study ML via MCEM in this subsection in order to study the connections between ML, MI, and FB, and to shed light on examining the properties of the MCEM method when closed form estimates under ML do not exist. ML via MCEM will be the basis of our development in this subsection. In particular, we will derive expressions for the estimates, and their associated variances for both the small sample and large sample situations using MCEM. Following [6] and [8], the Monte Carlo E-Step at the $(t+1)^{st}$ EM iteration can be written as

$$Q(\gamma|\gamma^{(t)}) = \int l(\gamma|\mathbf{D}_{obs}, \mathbf{D}_{mis}, \gamma^{(t)}) p(\mathbf{D}_{mis}|\mathbf{D}_{obs}, \gamma^{(t)})\, d\mathbf{D}_{mis} \approx \frac{1}{m}\sum_{j=1}^{m} l(\gamma|\mathbf{D}_{obs}, \mathbf{D}^{(j)}_{mis}, \gamma^{(t)}),$$

where $l(\gamma|\mathbf{D}_{obs}, \mathbf{D}_{mis}, \gamma^{(t)})$ is the log-likelihood function based on the complete data given the parameter estimates at the t^{th} iteration, $\mathbf{D}_{obs} = (\mathbf{y}_1, \mathbf{X}_1, \mathbf{X}_2)$ is the observed data, $\mathbf{D}_{mis} = \mathbf{y}_2$ and the $\mathbf{D}_{mis}^{(j)}$'s are the missing values replaced by their j^{th} sampled values from the full conditional distribution $p(\mathbf{D}_{mis}|\mathbf{D}_{obs}, \gamma^{(t)})$. The M-Step at the $(t+1)^{st}$ EM iteration maximizes $Q(\gamma|\gamma^{(t)})$. Standard errors can be calculated by using Louis's method and the estimated observed information matrix of γ based on Louis's method is given by

$$I(\hat{\gamma}) = -\ddot{Q}(\hat{\gamma}|\hat{\gamma}) - \frac{1}{m}\sum_{j=1}^{m} S(\hat{\gamma};\mathbf{D}_{obs}, \mathbf{D}_{mis}^{(j)}) S(\hat{\gamma};\mathbf{D}_{obs}, \mathbf{D}_{mis}^{(j)})'$$

where $\hat{\gamma}$ is the ML estimate at MCEM convergence and $\ddot{Q}(\hat{\gamma}|\hat{\gamma})$ is the second derivative matrix of the Q function. The estimate of the asymptotic covariance matrix of $\hat{\gamma}$ is therefore $[\mathcal{I}(\hat{\gamma})]^{-1}$.

Note that unlike the MI method, which creates m pseudo complete datasets by replacing the missing values with each of the m sets of imputed values, ML via MCEM calculates the estimates from a single dataset and assigns a weight of 1 for complete observations and a weight of $1/m$ for each sampled value. In order to explore the connections between MI and ML, we consider the imputation distribution $[\mathbf{y}_2|\mathbf{y}_1, \hat{\beta}]$ of MCEM, given by

$$\mathbf{y}_2|\mathbf{y}_1, \hat{\beta} \sim N_{n-n_1}\left(\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1, \left(\frac{n_1-p}{n_1}\right)s^2\mathbf{I}\right), \quad (14)$$

where $s^2 = \mathbf{y}_1'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}_1/(n_1-p)$. Theorem 2.2 gives the estimates of β and σ^2 along with their small and large sample properties.

Theorem 2.2—For the linear regression model (1), let $\mathbf{y}_2^{(l)}$, $l = 1, \ldots, m$, be the Gibbs samples of $\mathbf{y}_2$ from $[\mathbf{y}_2|\mathbf{y}_1, \hat{\beta}]$ in Eq. (14). Then

i. the maximum likelihood estimate of β using MCEM is

$$\hat{\beta}_{ML2} = \frac{1}{m}\sum_{l=1}^{m}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_2^{(l)}), \quad (15)$$

with variance

$$Var(\hat{\beta}_{ML2}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} + \frac{\sigma^2(n_1-p)}{mn_1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}'\mathbf{X})^{-1}. \quad (16)$$

ii. $\hat{\beta}_{ML2}$ is an unbiased estimator of β.

iii. The ML estimate of σ^2 is

$$\hat{\sigma}^2_{ML2}=\frac{1}{m}\sum_{l=1}^{m}(\mathbf{y}^{(l)}-\mathbf{X}\hat{\beta}_{ML2})'(\mathbf{y}^{(l)}-\mathbf{X}\hat{\beta}_{ML2})/n \quad (17)$$

with mean

$$E(\hat{\sigma}^2_{ML2})=\left(\frac{n_1-p}{n_1}-\frac{(n_1-p)tr(M)}{mnn_1}\right)\sigma^2$$

and variance

$$Var(\hat{\sigma}^2_{ML2})=\frac{2(n_1-p)}{n_1^2}\sigma^4+\frac{2(n_1-p)}{n^2n_1^2}a_2\sigma^4, \quad (18)$$

where $a_2=\frac{tr^2(M)+(n_1-p+2)\,tr(M^2)-2(n_1-p+2)tr(M)}{m^2}+\frac{(n-n_1)(n_1-p+2)-2n\,tr(M)}{m}$ and $M=(X_2'X_2)\,(X'X)^{-1}$.

iv. $\hat{\sigma}^2_{ML2}\xrightarrow{\mathscr{P}}\sigma^2$, as $n_1\to\infty$ and $m\to\infty$.

Again from Theorem 2.2, it can be easily shown that the estimate of β and its variance based on MCEM are asymptotically equivalent to the CC estimates. In particular, after some algebra, it can be shown that

$$\frac{E(\hat{\sigma}^2_{ML2}|\mathbf{y}_1)}{\hat{\sigma}^2_{CC}}=1-\frac{\mathrm{tr}(M)}{mn}\to 1,$$

as n_1 or $m\to\infty$. The condition that tr(M) $\to K$, $0\le K<\infty$, as $n\to\infty$, implies that the information contained in the covariates corresponding to the missing responses is finite compared to the total information in the covariates. The variance of $\hat{\sigma}^2_{ML2}$ in Eq. (18) can also be written as

$$\mathrm{Var}(\hat{\sigma}^2_{ML2})=\mathrm{Var}(\hat{\sigma}^2_{CC})+O\left(\frac{1}{mn_1}\right)\sigma^4,$$

and hence $\mathrm{Var}(\hat{\sigma}^2_{ML2})/\mathrm{Var}(\hat{\sigma}^2_{CC})\to 1$ as n_1 or m go to infinity.

Note that the variance of $\hat{\beta_{ML2}}$ in Eq. (16) is smaller than the variance of $\hat{\beta_{MI}}$ in Eq. (11), however, the derivation of Theorem 2.2 is based on the assumption that the imputation distribution of the missing responses yields the ML estimates, which may not be true in practice. Again, note that although we write the estimates of (β, σ^2) and their variance as if there were m data sets in order to compare the MI and ML methods, in practice, ML via MCEM calculates the estimates from only one dataset with different weights assigned to the

observed and sampled values. In this sense, MCEM augments the data "vertically" and MI augments the data "horizontally".

Remark 2.3—Both $E(\hat{\sigma}^2_{ML2})$ and $\mathrm{Var}(\hat{\sigma}^2_{ML2})$ are functions of *m*, the number of Gibbs samples, and therefore, increasing *m* reduces the bias and variance of $\hat{\sigma}^2_{ML2}$.

2.3 Fully Bayesian (FB)

FB methods for the missing data problem are based on specifying priors for all of the parameters and then the missing data are sampled from their full conditional distributions within the Gibbs sampler. Clearly, ML and MI have Bayesian connections, since ML can be viewed as a large sample Bayesian method, and in many cases, the implementation of Bayesian methods using uniform improper priors on all parameters leads to ML estimates. In this subsection, we consider the FB method under conjugate priors, which yield closed form expressions for the posterior mean and variance of the parameters.

Note that observed data likelihood for MAR responses is the CC likelihood and thus the posterior distribution of γ^* based on the observed data is $p(\gamma^*|\mathbf{y}_1, \mathbf{X}) \propto p(\mathbf{y}_1|\mathbf{X}; \gamma^*)\pi(\gamma^*)$. Theorem 2.3 provides the properties of the fully Bayesian estimates of β and τ.

Theorem 2.3—For the linear regression model (1), assume that the prior for $\gamma^* = (\beta, \tau)$ is $\pi(\gamma^*) = \pi(\beta|\tau)\pi(\tau)$, where $\pi(\beta|\tau) = N(\mu_0, \tau^{-1}\mathbf{\Sigma_0})$ and $\pi(\tau) = \mathrm{Gamma}(\delta_0/2, \lambda_0/2)$. Then

i. the fully Bayesian estimate of β is

$$\hat{\beta}_{FB} = \frac{1}{m}\sum_{l=1}^{m}\beta^{(m)},$$

where β_m is the *m* sample from the posterior distribution

$$p(\beta|D_{obs}) \sim S_p(n_1+\delta_0, \tilde{\beta}, \tilde{s}^2((\mathbf{X}_1'\mathbf{X}_1+\textstyle\sum_0)^{-1}))$$

with $\tilde{\beta} = \Lambda\mu_0 + (\mathbf{I} - \mathbf{\Lambda})\hat{\beta}$, $\Lambda = (\mathbf{X}_1'\mathbf{X}_1+\sum_0^{-1})^{-1}\sum_0^{-1}$, $\hat{\beta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$, and $\tilde{s}^2 = (n_1+\delta_0)^{-1}[\mathbf{y}_1'(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}_1 + (\hat{\beta}-\mu_0)'(\mathbf{\Lambda}'\mathbf{X}_1'\mathbf{X}_1)(\hat{\beta}-\mu_0) + \lambda_0]$.

ii. The posterior mean and variance of β based on the observed data are

$$E(\beta|\mathbf{D}_{obs}) = \tilde{\beta}$$

and

$$Var(\beta|\mathbf{D}_{obs}) = (n_1+\delta_0)\tilde{s}^2(\mathbf{X}_1'\mathbf{X}_1+\textstyle\sum_0)^{-1}/(n_1+\delta_0-2).$$

iii. The fully Bayesian estimate of $\tau = 1/\sigma^2$ is

$$\hat{\tau}_{FB}=\frac{1}{m}\sum_{l=1}^{m}\tau^{(m)}$$

where $p(\tau|D_{obs}) \sim \text{Gamma}((n_1+\delta_0)/2, (n_1+\delta_0)s^{2\tilde{}}/2)$ with $s^{2\tilde{}}$ defined in (**i**).

iv. The posterior mean and variance of τ are

$$E(\tau|\mathbf{D}_{obs})=1/s^2 \;\; and \;\; Var(\tau|\mathbf{D}_{obs})=2(n_1+\delta_0)^{-1}\tilde{s}^{-4}.$$

The proof of Theorem 2.3 is straightforward and can be found in most Bayesian textbooks. We state it as a Theorem here only to be consistent with other sections.

Remark 2.4—When the prior for γ^* is an improper prior, $\pi(\gamma^*) \propto \tau^{-1}$, $s^{2\tilde{}}$ reduces to $\mathbf{y}_1'(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}_1/n_1$ and the posterior mean and variance of (β, τ) are then equal to the CC estimates given in equations (2) and (3).

Therefore, the CC analysis is recommended over the MI, MCEM, and FB methods for MAR responses in the linear regression model, unless additional information is available to specify informative priors for the MI and FB methods, or the imputation model of MI includes covariates not specified in the analysis model. On the other hand, the loss of efficiency of MI, MCEM, and FB methods can be significantly reduced by increasing the number of imputations for MI or the number of Gibbs samples for MCEM and FB.

3. SIMULATION STUDY

In this section, we will compare inferences about β using the four methods, MI, CC, MCEM and FB using the formulas we developed in Section 2 for a small sample size n and various values of m for MI and MCEM.

We generate N = 1,000 replicates with each simulation consisting of n = 250 independent response variables y_i from the linear regression model as

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+e_i,$$

where the e_i's are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$. The values chosen for the parameters are $(\beta_0, \beta_1, \beta_2) = (1.0, 1.5, -1.0)$ and $\sigma^2 = 1.0$. The covariates (x_{i1}, x_{i2}) are i.i.d. and simulated as

$$x_{i1}\sim N(1.0, 1.0) \quad \text{and} \quad x_{i2}|x_{i1}\sim N(\alpha_0+\alpha_1 x_{i1}, \sigma_x^2)$$

where $(\alpha_0, \alpha_1) = (1.0, 1.0)$ and $\sigma_x^2=1.0$.

We assume that y_i is MAR for some i's and x_{i1} and x_{i2} are completely observed throughout. In this setting, the model for the missing data mechanism of y_i is given by

$$p(r_{i1}=1|x_{i1}, x_{i2}, \phi)=\frac{\exp(\phi_0+\phi_1 x_{i1}+\phi_2 x_{i2})}{1+\exp(\phi_0+\phi_1 x_{i1}+\phi_2 x_{i2})},$$

where $(\varphi_{01}, \varphi_{11}, \varphi_{21}) = (-5.5, 1.0, 1.0)$ and $r_{i1} = 1$ if y_i is missing, 0 otherwise.

Table 1 gives the results using the four methods, MI, CC, MCEM and FB, and also gives the estimates based on the full data (i.e., no missing values), as these estimates serve as a benchmark for comparison. From the $N = 1{,}000$ simulations, the average number of observations with y_i missing is 19%. We chose the number of samples m equal to 30 and 3 in both the MI and MCEM methods in order to compare the results. [13] note that for proper MI, m, the number of imputed datasets, can be as small as $m = 5$. However, m in MCEM, the number of Gibbs samples, is usually large, say $m = 100$ or more, in order to accurately represent the sampling distribution in the E-step, especially in complex models with large missing data fractions. This is consistent with the simulation results. When $m = 3$ in the MI method, the two forms of the variance estimates give similar coverage rates because both of them adjust well for small m, and when $m = 3$ in MCEM, equations (16) and (18) give much better coverage rates than the Louis method. On the other hand, for the MI method, considering that the estimates with $m = 30$ always have smaller variances than the estimates with $m = 3$ with better coverage probabilities, larger values of m may need to be considered if the computational burden is not heavy. The simulation results confirm the theorems in Section 2, and show that the three methods (MI, ML via MCEM, FB) produce consistent estimates with valid inferences and all are asymptotically equivalent to the CC estimates when the response variable is MAR.

4. LIVER CANCER DATA

To further illustrate the CC, MI, ML and FB methods, we consider a real dataset on $n = 174$ patients from two Eastern Cooperative Oncology Group clinical trials, EST 2282 [3] and EST 1286 [4]. We are interested in how the number of cancerous liver nodes (CNT) when entering the trials is predicted by six other baseline characteristics: body mass index (BMI, defined as weight in kilograms divided by the square of height in meters); age (in years); associated jaundice (yes, no); and time since diagnosis of the disease (TSD, in weeks). Thirty four out of 174 (19.5%) patients have a missing response variable (CNT). Throughout, we assume that the response variable CNT is MAR. The square root transformation on CNT and TSD was used in the analyses.

We use linear regression to model the response variable, $\sqrt{CNT}$, as $\sqrt{CNT_i}=\beta_0+\beta_1 \mathrm{BMI}_i+\beta_2 \mathrm{Age}_i+\beta_3 \mathrm{Jaundice}_i+\beta_4 \sqrt{TSD_i}+e_i$, where the e_i's are i.i.d. normally distributed as $e_i \sim N(0, \sigma^2)$.

Table 2 show the results for the CC analysis, MI with $m = 3$ and $m = 30$ using equations (11) and (13), MCEM with $m = 30$ and $m = 300$ using equations (16) and (18), and the FB

method discussed in Section 3. Moreover, the variance estimates are 0.775, 0.715, 0.715, 0.724, 0.739, and 0.712 for CC, MI with $m = 3$ and $m = 30$, ML via MCEM with $m = 30$ and $m = 300$, and FB, respectively. As shown in the table, the MI, MCEM and FB methods yield very similar estimates with very little differences from the CC estimates. In particular, the p-values of all the covariates except Age are smaller with larger m. The results show that the age of the patients is significantly associated with the number of cancerous liver nodes controlling for body mass index, associated jaundice, and time since diagnosis.

5. DISCUSSION

It is known in the missing data literature that when the responses are MAR, the CC analysis is unbiased and efficient. However, MI, ML via MCEM, and FB are still commonly used in practice in this setting. This may be due to the fact that (a) the unbiasedness and efficiency properties of the CC method in this setting is not known to general researchers; (b) MI, ML, and FB, as well as some other methods including parametric fractional imputation [10] and empirical likelihood [18] outperform CC in a general setting. To overcome these barriers, it is important to inform researchers and practitioners about these important results. Moreover, we also showed in this paper that the loss of efficiency of MI, ML via MCEM, and FB can be significantly reduced by increasing the number of imputations for MI and the number of Gibbs samples for MCEM and FB. It would be of interest to extend our theoretical results to MAR responses for models other than linear regression. This is a topic of current investigation. It would also be interesting to accommodate missingness in the predictors. Unfortunately, even for linear regression models with normally distributed MAR covariates, no closed form expressions are available for the estimates of the three methods, which makes the comparisons between the methods very hard. A special scenario of it, assuming unit variances for response variable and missing covariates, was investigated in [2] for the maximum likelihood approach.

Acknowledgments

The authors wish to thank the editor, the associate editor and two referees for several suggestions and editorial changes which have greatly improved the paper.

References

1. Barnard J, Rubin D. Small-sample degrees of freedom with multiple imputation. Biometrika. 1999; 86:948–955.
2. Chen Q, Ibrahim JG, Chen MH, Senchaudhuri P. Theory and inference for regression models with missing response and covariates. Journal of Multivariate Analysis. 2008; 99:1302–1331. [PubMed: 19169388]
3. Falkson G, Cnaan A, Simson IW. A randomized phase II study of acivicin and déoxydoxorubicin in patients with hepatocellular carcinoma in an eastern cooperative oncology group study. American Journal of Clinical Oncology. 1990; 13:510– 515. [PubMed: 2173394]
4. Falkson G, Lipsitz S, Borden E, Simson IW, Haller D. A ECOG randomized phase i1 study of beta interferon and menogoril. American Journal of Clinical Oncology. 1995; 18:287–292. [PubMed: 7625367]
5. Graybill, FA. Matrices with Applications in Statistics. 2. Duxbury Press; 1983. p. MR0682581Wadsworth Statistics/Probability Series

6. Ibrahim JG. Incomplete data in generalized linear models. Journal of the American Statistical Association. 1990; 85:765–769.
7. Ibrahim JG, Chen MH, Lipsitz SR. Bayesian methods for generalized linear models with covariates missing at random. Canadian Journal of Statistics. 2002; 30:55–78.
8. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linearmodels when the missing data mechanism is nonignorable. Journal of the Royal Statistical Society: Series B. 1999; 61:173–190.
9. Kim JK. Finite sample properties of multiple imputation estimators. The Annals of Statistics. 2004; 32:766–783.
10. Kim JK. Parametric fractional imputation for missing data analysis. Biometrika. 2011; 98:119–132.
11. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika. 1996; 83:916–922.
12. Little RJA. Inference about means from incomplete multivariate data. Biometrika. 1976; 63:593–604.
13. Little, RJA.; Rubin, DB. Statistical Analysis With Missing Data. 2. John Wiley; 2002. p. MR1925014
14. Louis T. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B. 1982; 44:226–233.
15. Meng XL, Romero M. Discussion to S. F. Nielsen: Efficiency and self-efficiency with multiple imputation inference. International Statistical Review. 2003; 71:607–618.
16. Meng XL, Rubin DB. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association. 1991; 86:899–909.
17. Nielsen SF. Proper and improper multiple imputation. International Statistical Review. 2003; 71:593–607.
18. Qin J, Zhang B, Leung DHY. Empirical likelihood in missing data problems. Journal of the American Statistical Association. 1999; 104:1492–1503.
19. Robins JM, Wang N. Inference for imputation estimators. Biometrika. 2000; 87:113–124.
20. Rubin D. Discussion to S. F. Nielsen: Discussion on multiple imputation. International Statistical Review. 2003; 71:619–625.
21. Schenker N, Welsh AH. Asymptotic results for multiple imputation. The Annals of Statistics. 1988; 16:1550–1566.
22. Wang N, Robins JM. Large-sample theory for parametric imputation procedures. Biometrika. 1998; 85:935–948.

APPENDIX

Proof of Lemma 2.1

If the $n \times 1$ random vector $\mathbf{z}$ has a multivariate t distribution as $S_n(v, \mu, \mathbf{V})$, then we can write $\mathbf{z}=\mathbf{x}/\sqrt{y/v}+\mu$, where $\mathbf{x}$ is an $n \times 1$ random vector that has a multivariate normal distribution $N(\mathbf{0}, \mathbf{V})$, y is a random variable which has a χ^2_v distribution, $\mathbf{x}$ and y are independent. Therefore, (i) is straightforward and, to get (ii), we have

$$\begin{aligned} E(\mathbf{z}\mathbf{z}'\mathbf{A}\mathbf{z}) &= E\left[E\left(\frac{\mathbf{x}\mathbf{x}'\mathbf{A}\mathbf{x}}{(y/v)^{3/2}}+\frac{\mathbf{x}\mu'\mathbf{A}\mathbf{x}+\mathbf{x}\mathbf{x}'\mathbf{A}\mu+\mu\mathbf{x}'\mathbf{A}\mathbf{x}}{(y/v)}+\frac{\mu\mu'\mathbf{A}\mathbf{x}+\mu\mathbf{x}'\mathbf{A}\mu}{(y/v)^{1/2}}+\mu\mu'\mathbf{A}\mu|\mathbf{y}\right)\right] \\ &= \mathbf{0}+E\left[vy^{-1}(\mathbf{V}\mathbf{A}'\mu+\mathbf{V}\mathbf{A}\mu+\mu\mathrm{tr}(\mathbf{A}\mathbf{V}))\right]+\mathbf{0}+\mu\mu'\mathbf{A}\mu \\ &= \frac{v}{v-2}[\mathbf{V}\mathbf{A}'\mu+\mathbf{V}\mathbf{A}\mu+\mu\mathrm{tr}(\mathbf{A}\mathbf{V})]+\mu\mu'\mathbf{A}\mu. \end{aligned}$$

We substitute $\mathbf{z}=\mathbf{x}/\sqrt{y/v}+\mu$ and calculate the double expectation in the first equality. Because of independence between **x** and *y*, we can substitute in expressions for multivariate normal moments in the second equality, and therefore

$$\begin{aligned}E[(\mathbf{z}'\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{B}\mathbf{z})] &= E[E((\mathbf{x}/\sqrt{y/v}+\mu)'\mathbf{A}(\mathbf{x}/\sqrt{\mathbf{y}/\mathbf{v}}+\mu)(\mathbf{x}/\sqrt{\mathbf{y}/\mathbf{v}}+\mu)' \times \mathbf{B}(\mathbf{x}/\sqrt{y/v}+\mu)|\mathbf{y})] \\ &= E[(v/y)^2 E(\mathbf{x}'\mathbf{A}\mathbf{x}\mathbf{x}'\mathbf{B}\mathbf{x})] + E[(v/y)^{3/2} \times \mathbf{0}] \\ &+ E[(v/y)E(\mathbf{x}'\mathbf{A}\mathbf{x}\mu'\mathbf{B}\mu + \mu'\mathbf{A}\mathbf{x}\mu'\mathbf{B}\mathbf{x} + \mu'\mathbf{A}\mathbf{x}\mathbf{x}'\mathbf{B}\mu \\ &+ \mu'\mathbf{A}\mathbf{x}\mathbf{x}'\mathbf{B}\mu + \mathbf{x}'\mathbf{A}\mu\mu'\mathbf{B}\mathbf{x} + \mathbf{x}'\mathbf{A}\mu\mathbf{x}'\mathbf{B}\mu \\ &+ \mu'\mathbf{A}\mu\mathbf{x}'\mathbf{B}\mathbf{x})] + \mathbf{E}[(\mathbf{v}/\mathbf{y})^{\mathbf{1/2}} \times \mathbf{0}] + \mu'\mathbf{A}\mu\mu'\mathbf{B}\mu \\ &= \frac{v^2}{(v-2)(v-4)}[\mathrm{tr}(\mathbf{A}\mathbf{V})tr(\mathbf{B}\mathbf{V}) + \mathrm{tr}(\mathbf{A}\mathbf{V}\mathbf{B}\mathbf{V}) \\ &+ \mathrm{tr}(\mathbf{A}\mathbf{V}\mathbf{B}'\mathbf{V})] + \frac{v}{(v-2}[\mathrm{tr}(\mathbf{A}\mathbf{V})\mu'\mathbf{B}\mu + \mu'\mathbf{A}\mu\mathrm{tr}(\mathbf{B}\mathbf{V}) \\ &+ \mu'(\mathbf{A}\mathbf{V}\mathbf{B} + \mathbf{A}'\mathbf{V}\mathbf{B} + \mathbf{A}\mathbf{V}\mathbf{B}' + \mathbf{A}'\mathbf{V}\mathbf{B}')\mu] \\ &+ \mu'\mathbf{A}\mu\mu'\mathbf{B}\mu.\end{aligned}$$

In the second equality, the two zero components correspond to the first and third moments of **x**. In the last equality, we use the second and fourth moments of **x**, which are available in [5] with a modification for nonsymmetric matrices **A** and **B**.

Proof of Theorem 2.1

(i) and (ii): It is straightforward to get the estimate of β as in Eq. (10). It is also straightforward to use double expectations to get $E(\hat{\beta}_{MI}) = \beta$. To find the variance of $\hat{\beta}_{MI}$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{MI}) &= E(\mathrm{Var}(\hat{\beta}|\mathbf{y}_1)) + \mathrm{Var}(E(\hat{\beta}|\mathbf{y}_1)) \\ &= \frac{n_1-p}{m(n_1-p-2)} \times E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2' s^2(\mathbf{I}+\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2')\mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-1}) + \mathrm{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{y}_1 + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1) \\ &= \frac{\sigma^2(n_1-p)}{m(n_1-p-2)}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1)^{-1} + \mathrm{Var}((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1) \\ &= \frac{\sigma^2(n_1-p)}{m(n_1-p-2)}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1)^{-1} + \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.\end{aligned}$$

(iii) and (iv): It is straightforward to get the estimate of σ^2 as in Eq. (12). In order to find $E(\hat{\sigma}^2_{MI})$, we write $(n-2)\hat{\sigma}^2_{MI} = \mathbf{y}_1'\mathbf{A}\mathbf{y}_1 - 2\mathbf{y}_1'\mathbf{B}\bar{\mathbf{y}}_2 + \frac{1}{m}\sum_{l=1}^{m}\mathbf{y}_2^{(l)'}\mathbf{C}\mathbf{y}_2^{(l)}$, where $\bar{\mathbf{y}}_2 = \sum_{l=1}^{m}\mathbf{y}_2^{(l)}/m$, $\mathbf{A} = \mathbf{I} - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'$, $\mathbf{B} = \mathbf{X}_1(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}_2'$, and $\mathbf{C} = \mathbf{I} - \mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2'$. Let $\mathbf{P}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{D} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'$. Then after some algebra, we have

$$\begin{aligned} E(\hat{\sigma}^2_{MI}) &= E\left[E(\sigma^2_{MI}|\mathbf{y}_1)\right] \\ &= E\left[\mathbf{y}_1'(\mathbf{A}-\mathbf{D}+\tfrac{(n-n_1)}{(n_1-p-2)}\mathbf{P}_{\mathbf{x}_1})\mathbf{y}_1\right]/(n-2) \\ &= \tfrac{(n-p-2)}{(n_1-p-2)\,(n-2)}E(\mathbf{y}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1) \\ &= \tfrac{(n-p-2)\,(n_1-p)}{(n-2)\,(n_1-p-2)}\sigma^2 \\ &\to \sigma^2, \end{aligned}$$

by noting that $\mathbf{y}_1 \sim N(\mathbf{X}_1\beta, \sigma^2\mathbf{I})$, $\mathbf{A} - \mathbf{D} = \mathbf{P}_{\mathbf{X}_1}$, $\mathbf{P}^2_{\mathbf{x}_1} = \mathbf{P}_{\mathbf{x}_1}$ and $\mathbf{X}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{X}_1 = \mathbf{0}$.

To find $\mathrm{Var}(\hat{\sigma}^2_{MI})$, we write

$\mathrm{Var}((n-2)\hat{\sigma}^2_{MI}) = \mathrm{Var}\left[E((n-2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)\right] + E\left[\mathrm{Var}((n-2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)\right]$. First we have

$$\begin{aligned} \mathrm{Var}\left[E((n-2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)\right] &= \tfrac{(n-p-2)^2}{(n_1-p-2)^2}\mathrm{Var}((n_1-p)s^2) \\ &= \tfrac{2(n_1-p)\,(n-p-2)^2}{(n_1-p-2)^2}\sigma^4. \end{aligned}$$

Then we obtain

$$\begin{aligned} \mathrm{Var}((n-2)\hat{\sigma}^2_{MI}|\mathbf{y}_1) &= \mathrm{Var}(-2\mathbf{y}_1'\mathbf{B}\overline{\mathbf{y}}_2 + \tfrac{1}{m}\sum_{l=1}^{m}\mathbf{y}_2^{(l)'}\mathbf{C}\mathbf{y}_2^{(l)}|\mathbf{y}_1) \\ &= \tfrac{4\mathbf{y}_1'\mathbf{B}\mathrm{Var}(\mathbf{y}_2|\mathbf{y}_1)\mathbf{B}'\mathbf{y}_1}{m} + \tfrac{\mathrm{Var}(\mathbf{y}_2'\mathbf{C}\mathbf{y}_2|\mathbf{y}_1)}{m} - \sum_{l=1}^{m}\sum_{k=1}^{m}\tfrac{4\mathrm{Cov}(\mathbf{y}_1'\mathbf{B}\mathbf{y}_2^{(k)}, \mathbf{y}_2^{(l)}\mathbf{C}\mathbf{y}_2^{(l)}|\mathbf{y}_1)}{m^2} \\ &= \tfrac{4(n_1-p)}{m(n_1-p-2)}\mathbf{y}_1'\mathbf{B}\textstyle\sum\mathbf{B}'\mathbf{y}_1 + \tfrac{1}{m}\mathrm{Var}(\mathbf{y}_2'\mathbf{C}\mathbf{y}_2|\mathbf{y}_1) - \tfrac{4}{m}\mathrm{Cov}(\mathbf{y}_1'\mathbf{B}\mathbf{y}_2, \mathbf{y}_2\mathbf{C}\mathbf{y}_2|\mathbf{y}_1), \end{aligned}$$

where $\sum = s^2(\mathbf{I} + \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2')$ and s^2 is defined as in Theorem 2.1. The last equality holds because $\mathbf{y}_2^{(j)}$ is independent of $\mathbf{y}_2^{(k)}$ given $\mathbf{y}_1$, when $j \neq k$. Then we use Lemma 2.1 and get

$$E\left[\mathrm{Var}((n-2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)\right] = \frac{2(n_1-p)\,(n-n_1)\,(n-p-2)}{(n_1-p-2)^2(n_1-p-4)}E\left[\frac{\mathbf{y}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1\mathbf{y}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1}{m}\right] = \frac{2(n_1-p)\,(n-n_1)\,(n-p-2)\,(n_1-p+2)}{m(n_1-p-2)^2(n_1-p-4)}\sigma^4,$$

and therefore, we can get $\mathrm{Var}(\hat{\sigma}^2_{MI})$ as in (13). Since $\mathrm{Var}(\hat{\sigma}^2_{MI}) \to 0, \hat{\sigma}^2_{MI} \xrightarrow{\mathscr{P}} \sigma^2$.

Proof of Theorem 2.2

(i) and (ii): It is straightforward to get the estimate of β as in Eq. (15). It is straightforward to use double expectations to get $E(\hat{\beta}_{MI}) = \beta$. To find the variance of $\hat{\beta}_{ML2}$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{ML2}) &= E(\mathrm{Var}(\hat{\beta}|\mathbf{y}_1)) + \mathrm{Var}(E(\hat{\beta}|\mathbf{y}_1)) \\ &= \tfrac{n_1-p}{mn_1} E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2' s^2 \mathbf{I} \mathbf{X}_2 (\mathbf{X}'\mathbf{X})^{-1}) + \mathrm{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{y}_1 + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1) \\ &= \tfrac{\sigma^2(n_1-p)}{mn_1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}'\mathbf{X})^{-1} + \mathrm{Var}((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1) \\ &= \tfrac{\sigma^2(n_1-p)}{mn_1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.\end{aligned}$$

(iii) and (iv): It is straightforward to get the estimate of σ^2 as in Eq. (17). In order to find $E(\hat{\sigma}^2_{MI})$, we write $n\hat{\sigma}^2_{ML2} = \mathbf{y}_1'\mathbf{A}\mathbf{y}_1 + \sum_{l=1}^{m} \mathbf{y}_2^{(l)'}\mathbf{y}_2^{(l)}/m - 2\mathbf{y}_1'\mathbf{B}\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_2^{(l)'}(\mathbf{I}-\mathbf{C})\bar{\mathbf{y}}_2^{(l)}$, where the symbols are same as the proof of Theorem 2.2. Then we have

$$\begin{aligned}E(\hat{\sigma}^2_{ML2}) &= E\left[E(\hat{\sigma}^2_{MI}|\mathbf{y}_1)\right] \\ &= E\left[\mathbf{y}_1'(\mathbf{A}-\mathbf{D}+\tfrac{n-n_1}{n_1}\mathbf{P}_{\mathbf{x}_1} - \tfrac{\mathrm{tr}(\mathbf{M})}{mn_1}\mathbf{P}_{\mathbf{x}_1})\mathbf{y}_1\right]/n \\ &= \left(\tfrac{n_1-p}{n_1} - \tfrac{(n_1-p)\mathrm{tr}(M)}{mnn_1}\right)\sigma^2 \\ &\to \sigma^2,\end{aligned}$$

where $M = (X_2' X_2)(X'X)^{-1}$. To find $\mathrm{Var}(\hat{\sigma}^2_{ML2})$, we have

$$\mathrm{Var}((n+2)\hat{\sigma}^2_{MI}) = \mathrm{Var}[E((n+2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)] + E[\mathrm{Var}((n+2)\hat{\sigma}^2_{MI}|\mathbf{y}_1)].$$

Then we have

$$\begin{aligned}\mathrm{Var}[E(n\hat{\sigma}^2_{ML2}|\mathbf{y}_1)] &= \mathrm{Var}\left(\left(\tfrac{n}{n_1} - \tfrac{\mathrm{tr}(M)}{mn}\right)\mathbf{y}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1\right) \\ &= 2\left(\tfrac{n}{n_1} - \tfrac{\mathrm{tr}(M)}{mn}\right)^2 (n_1-p)\sigma^4.\end{aligned}$$

Using Chapter 10.9 of [5] and after some algebra, we get

$$\mathrm{Var}(n\hat{\sigma}^2_{ML2}|\mathbf{y}_1) = 2\left(\frac{n-n_1}{mn_1^2} - \frac{\mathrm{tr}(\mathbf{M}^2)}{m^2n_1^2} - \frac{2\mathrm{tr}(\mathbf{M})}{m^2n_1^2}\right)(\mathbf{y}_1'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1)^2.$$

Therefore, after some algebra, we can get $\mathrm{Var}(\hat{\sigma}^2_{ML2})$ as in Eq. (18), and therefore $\hat{\sigma}^2_{ML2} \xrightarrow{\mathscr{P}} \sigma^2$.

Table 1

Simulation with MAR responses in the linear regression model. The 95% CR is the coverage rate of a 95% confidence interval. $\hat{\gamma_F}$ (full data), $\hat{\gamma}^*_{MI}$ (Multiple Imputation with covariance matrix as Eq. (7), $\hat{\gamma}^\dagger_{MI}$ (Multiple Imputation with covariance matrix as Eq. (11) and Eq. (13), $\hat{\gamma_{CC}}$ (CC estimates), $\hat{\gamma}^*_{ML2}$ (MCEM with covariance matrix using Louis's method, $\hat{\gamma}^\dagger_{ML2}$ (MCEM with covariance matrix as Eq. (16) and Eq. (18), and $\hat{\gamma_{FB}}$ (fully Bayesian). The number *m* is defined in the section discussing the corresponding method

Method	***m***	***Estimate (var(×10^{-3}))[95% CR]***			
		β_0 = 1.00	**β_1 = 1.50**	**β_2 = −1.00**	**σ^2 = 1.0**
$\hat{\gamma_F}$	-	0.994(12)[94]	1.499(8)[96]	−0.998(4)[96]	0.989(8)[94]
$\hat{\gamma}^*_{MI}$	30	0.995(14)[95]	1.496(11)[95]	−0.998(5)[95]	1.000(10)[95]
$\hat{\gamma}^\dagger_{MI}$	30	0.995(14)[95]	1.496(10)[95]	−0.998(5)[95]	1.000(10)[95]
$\hat{\gamma}^*_{MI}$	3	0.993(14)[95]	1.496(11)[95]	−0.998(6)[94]	1.002(11)[94]
$\hat{\gamma}^\dagger_{MI}$	3	0.993(14)[95]	1.496(11)[96]	−0.998(6)[95]	1.002(10)[94]
$\hat{\gamma_{CC}}$	-	0.995(14)[95]	1.497(10)[95]	−0.9984(5)[95]	0.987(10)[94]
$\hat{\gamma}^*_{ML2}$	30	0.995(14)[95]	1.497(10)[95]	−0.998(5)[95]	0.987(10)[94]
$\hat{\gamma}^\dagger_{ML2}$	30	0.995(14)[95]	1.497(10)[95]	−0.998(5)[95]	0.987(10)[94]
$\hat{\gamma}^*_{ML2}$	3	0.993(14)[92]	1.497(10)[90]	−0.998(5)[89]	0.987(10)[90]
$\hat{\gamma}^\dagger_{ML2}$	3	0.993(14)[95]	1.497(10)[95]	−0.998(5)[93]	0.987(10)[93]
$\hat{\gamma_{FB}}$	-	0.995(14)[95]	1.497(10)[95]	−0.998(5)[95]	0.992(10)[95]

Table 2

Estimates for liver cancer data

Effect	Method	$\hat{\beta}$	Std	P-value
Intercept	CC	2.826	0.445	< .001
	MI $m = 3$	2.801	0.446	< .001
	MI $m = 30$	2.763	0.408	< .001
	ML2 $m = 30$	2.782	0.440	< .001
	ML2 $m = 300$	2.766	0.395	< .001
	FB	2.770	0.408	< .001
BMI	CC	−0.008	0.015	0.595
	MI $m = 3$	−0.007	0.015	0.632
	MI $m = 30$	−0.005	0.013	0.726
	ML2 $m = 30$	−0.005	0.014	0.740
	ML2 $m = 300$	−0.005	0.013	0.715
	FB	−0.005	0.013	0.720
Age	CC	−0.012	0.005	0.018
	MI $m = 3$	−0.012	0.005	0.016
	MI $m = 30$	−0.012	0.005	0.007
	ML2 $m = 30$	−0.012	0.005	0.012
	ML2 $m = 300$	−0.012	0.004	0.006
	FB	−0.012	0.005	0.008
Jaundice	CC	0.190	0.152	0.212
	MI $m = 3$	0.231	0.146	0.115
	MI $m = 30$	0.217	0.134	0.107
	ML2 $m = 30$	0.210	0.144	0.147
	ML2 $m = 300$	0.204	0.129	0.116
	FB	0.204	0.134	0.129
$\sqrt{TSD}$	CC	0.002	0.034	0.964
	MI $m = 3$	0.000	0.034	0.998
	MI $m = 30$	−0.002	0.032	0.957
	ML2 $m = 30$	−0.003	0.034	0.933
	ML2 $m = 300$	−0.003	0.030	0.926
	FB	−0.002	0.032	0.943